DOCUMENT RESUME

ED 217 073 TM 820 326

AUTHOR McArthur, David L.
TITLE Bias in the Writing of Prose and Its Appraisal.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO CSE-RR-162
PUB DATE 81
NOTE 21p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; *Essays; *Ethnic Bias; Grade 5; Grade 6; Hispanic Americans; Intermediate Grades; *Minority Groups; Prose; Visual Stimuli; *Writing Evaluation
IDENTIFIERS *Inter Rater Reliability

ABSTRACT

Evidence from a variety of sources suggests that systematic differences can be found in the ratings given to student essays as a function not only of the student's skills but also of aspects of both the student's background and the background of the rater. Additionally, the nature of the prompt which provided the central theme of the essay might bias the outcome of the ratings of that essay. A study of ratings of fifth and sixth graders who wrote paragraph-long essays in response to two topics presented either in written or pictorial form is presented. Students were classified as Hispanic-surnamed or non-Hispanic-surnamed; two teachers, trained as raters using an objectively-based essay scoring scheme, represented an Hispanic cultural background and two a non-Hispanic background. Results from a blind rating of 100 complete essays show that several of the rating subscales were significantly influenced by an interaction between student ethnicity and rater ethnicity, and several subscales by rater ethnicity alone. Student ethnicity alone was not a significant main effect on any subscale. Prompt modality is significant for one subscale, and interacts with rater ethnicity on one other. The findings are interpreted as a direct indication of biased assessment. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED217073

BIAS IN THE WRITING OF PROSE AND ITS APPRAISAL

David L. McArthur

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM CS
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. Gray
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CSE Report No. 162
1981

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

TM 820 326

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Appreciation is extended to Edys Quellmalz and Frank Capell for their roles in the planning of this study, and to Chi Ping Chou and Beverly Cabello for their roles in the analysis.

ABSTRACT

Evidence from a variety of sources suggests that systematic differences can be found in the ratings given to student essays as a function not only of the student's skills but also of aspects of both the student's background and the background of the rater. Additionally, the nature of the prompt which provided the central theme of the essay might bias the outcome of the ratings of that essay. A study of ratings of fifth and sixth graders who wrote paragraph-long essays in response to two topics presented either in written or pictorial form is presented. Students were classified as Hispanic-surnamed or non-Hispanic-surnamed; two teachers, trained as raters using an objectively-based essay scoring scheme, represented an Hispanic cultural background and two a non-Hispanic background. Results from a blind rating of 100 complete essays show that several of the rating subscales were significantly influenced by an interaction between student ethnicity and rater ethnicity, and several subscales by rater ethnicity alone. Student ethnicity alone was not a significant main effect on any subscale. Prompt modality is significant for one subscale, and interacts with rater ethnicity on one other. The findings are interpreted as a direct indication of biased assessment.

Introduction

The evaluation of schoolchildren's prose writing poses special problems in relation to bias in educational appraisal. Many factors have long been known to have major influence on the prose writing performance of minority pupils. The literature on the issue of biases which occur in the judgement of students' written work is much smaller, and has proved much more contradictory. Are there specific aspects of non-native English writing style which undermine the usual procedures for judging writing performance? Do raters who match the cultural background of the writers whose work they judge arrive at different conclusions from raters who do not share the same background? In the present paper, the results of a research study involving both writers and readers from two different cultures are examined in an attempt to partition out the sources of systematic bias in the evaluation of writing.

Sources of Bias: Student Variables

An overarching concern in the literature about bias in writing has been the isolating of sociocultural factors in students' backgrounds which contribute to differences in performance. A half-century ago, Caldwell and Mowry (1933) demonstrated that bilingual Hispanic children, due to their use of language compared to their monolingual English-speaking counterparts, were at a disadvantage when evaluated by the essays they wrote; on objective examinations the differences were not nearly as acute. Parallel findings emerge from the recent large-scale study by White and Thomas (1981), who combined files of data regarding entering students in the California State University and Colleges system to yield graphic comparisons of total scores for 5,246 Whites, 585 Blacks, 449 Mexican-Americans, and 617 Asian-Americans on two English placement exams. The first was the CSUS's own English Placement Test; the second was the Test of Standard Written English from the College Entrance Examination Board. Although no statistical analyses were presented,

profiles of the four distributions suggest that a dialect interference or second language interference hurt the overall performance of the three minority samples on both tests. Lay (1978) has shown that native-speaking Chinese students are at a disadvantage in writing English prose because of the wide differences in structure and phonology of English and Chinese. Rizzo and Villafane (1978) have shown that similar explanation applies to native Spanish-speaking students.

Many investigators of language have shown that structural aspects of both oral and written language are significant in determining how children process the world around them. Moreover, many of the rules which govern functions of sending and receiving meaning using oral language are significantly different from those for written expression (Olson, 1977). For the non-native speaker of English the task of writing in English poses a particular problem because

> ...the surface structure of writing is an inadequate representation of both the sound structure of the target language and its meaning. Learning the underlying structure of the target language is as much of a bootstrap operation as the initial process of learning a mother tongue (Smith, 1975, p.359).

One practical outcome of such a structural viewpoint is that students who fail to acquire skills in the underlying structure of English might do passably well with spoken English but probably will have great difficulty with writing. Another factor not to be dismissed lightly is the attitudinal or psychological readiness of the student to orient positively to the task of acquiring skills in a new language (Cervantes, 1975; Lambert, Gardner, Barik, & Tunstall, 1963). Without the necessary motivation and appropriate learning context, students may be unable to let their knowledge of both the mother tongue and the new language interact to their advantage.

Sources of Bias: Evaluation Variables

Beyond the issues of students' involvement in languages lies an important

realm of educational and psychometric considerations having to do with the quantity and quality of appraisal. The nature of the task, how it is interpreted by both the student and the teacher, and the tools with which the students' writing is judged and by whom are all issues of import. In each of these lies the possibility of systematically different patterns of response for students from culturally or linguistically different groups. Each, then, may introduce its own bias into the evaluation of writing. The purpose of the writing task usually given to students in the classroom is to construct an essay following a particular prompt[1]. The teacher seeks a sufficient amount of this writing to rate the quality of the student's work. Exactly what elements are most important in that assessment of writing is often dependent upon the persons creating the scoring system. Freedman (1979) attempted to specify "definable parts" of student compositions which influenced teacher judgments. She concluded that content, organization, and language mechanics were the most important factors, in that order. The effect of "weak" content was so powerful that it overshadowed teacher judgment in every other category. The interaction of content quality judgments with the quality of the writing prompt is one point where bias in assessment is possible.

The use of incompletely explicated scoring criteria introduces another potential for bias in writing studies. In Rhodes-Hoover and Politzer's (1974) study of teachers' attitudes toward Black rhetoric, teachers downgraded compositions in the category of "language mechanics" because students failed to use

---

[1]The prompt itself may contribute to systematic bias. Some students may not know what the prompt represents because they do not completely understand the vocabulary of the prompt in written form, or do not recognize the pictorial content (the palm-tree vs. evergreen problem). Differences of an extreme nature are found in recognition of three-dimensional objects in photographs or drawings between children of developed and underdeveloped countries. Subtler problems of prompt recognizability abound: one British picture recognition test for the primary grades depicts electrical items common in England but totally unknown in America.

"superstandard" English. For example, if a student wrote, "I got there" as opposed to, "I reached my destination," the passage was considered too colloquial. Teachers not only gave their own interpretation of "usage" and "colloquial" but also imposed an undocumentable degree of severity in their judgment that may or may not have been intended by the scale.

In a study comparing the syntactical characteristics of Mexican and Anglo-American prose, Rodrigues (1978) asked educators whether they could detect "slight" or "noticeable" differences. More Anglo-American educators found "noticeable" differences than did Mexican-American raters. Bikson (1977) conducted a study of differences in working lexicons of 72 lower grade and 72 upper grade White, Chicano and Black elementary school students. Results showed that ethnically diverse speakers made <u>different</u> kinds of lexical choices, particularly in the early grades. The differences between Anglo lexicon and either the Black or Chicano lexicon were greater than the differences between the two minority lexicons. The study found varying degrees of overlap between minority and Anglo word choice. The minority students used a wider range of vocabulary than the Anglo group, but this "broader" working vocabulary is not often valued by persons evaluating the speech of these students.

Differences in classification of lexical terms between different linguistic groups may have consequences for the selection of scoring criteria to evaluate the writing of these groups. If we take concept classification tasks to be analogous to organization tasks in the writing process, then the different strategies used to associate words may reflect different preferred methods of essay organization. If the scoring criteria implicitly prefer one type of content organization strategy, such preference could result in bias against those students who adopt alternative strategies. Two studies in particular seem to suggest that words are

sorted by different ethnic groups into categories according to different classification strategies. Rissel (1978) studied the vocabulary-semantic relationship for monolingual English speakers, monolingual Spanish speakers and Spanish/English bilinguals living in New York and Puerto Rico to determine the classification strategies of these groups. The study found that not only did the classification strategies vary by linguistic group but that there appeared to be a relationship between amount of language dominance and classification strategy. Spanish dominant bilinguals employed comparative criteria, whereas the more "balanced" bilinguals used comparative classification for Spanish words and inclusive classification for English. Stahl (1977) conducted a study comparing the "methods for arrangement" of content used by Israeli students of European or Arabic extraction. He found that those of European background tended to arrange the content in a hierarchial or inclusive manner, whereas those of Arabic background tended to use more associative or comparative techniques. An interesting aspect of his method was that he gave higher points for hierarchial classification than for the use of comparative methods. In the assessment of writing this would appear to be deliberate introduction of biased criteria into the scoring process. Contrary results have been reported. In a study of syntactic patterns of lower and middle class Chicanos, Garcia (1975/76) concluded that the Chicanos used the same basic patterns found in American English, a conclusion also tendered by Rodrigues (1978). At the same time, however, Garcia cited research demonstrating differences in the morphological and phonological systems used by Chicanos and Anglos.

Recent informal evidence demonstrates the potency of systematic differences among raters of writing. Hartwell (1981) found that older, more experienced writers selected very different passages as exemplary of "professional writing" than did college freshmen. The differences appear to be consistent along a number of dimensions, including content, coherence, degree of complexity, and development.

Differences in rating of a written essay may also be related to the rater's own level of cognitive complexity and integration (Sternglass, 1981). Rater background has been found to influence how scoring criteria are interpreted and applied. Follman and Anderson (1967) concluded that when raters shared similar backgrounds with regard to education and opinions about what constitutes good writing, they tended to agree on the ratings of essays more than raters who differed along these dimensions.

Whether writing is assessed through normative-holistic means or through differentiated judgments on dimensions of rhetorical quality, the scoring "instrument" will always be a human judge. Consequently, no question about fairness, validity or accuracy in writing assessment can be fully addressed without reference to possible errors in judgment. The intention of writing assessment is to generate information useful for diagnosis and/or remediation. When diagnostic utility is of interest several other issues are pertinent. Diagnosis implies performance profiles which in turn require a multidimensional view of the writing skill domain. Questions about skill profiles are connected intimately to rater behavior in assigning ratings. Scoring criteria are filtered through the expectancies of raters, and the halo effect inflates inter-subscale correlations (Jaeger & Freijo, 1975). The use of more and longer writing tasks only exacerbates this phenomenon.

Rating scales may interact. It is common for writing score profiles to include some attention to essay "mechanics"; variations along this dimension may influence ratings on other dimensions. Ratings assigned to a writing sample on such dimensions as "organization" or "use of supporting detail" may be assigned differentially depending on the quality of mechanics within the essay. For mechanically substandard work, this process might bring the assessment of other

dimensions of writing quality into line with the rater's impression of mechanics, while if level of mechanics is not so low as to call attention to itself, there may be minimal confounding. However, across a given set of papers the net effect would be correlated true and error components and concomitant inflation of inter-subscale correlations. In a multitrait-multimethod factor analytic formulation the expectation in general would be for negative correlations between mechanics "trait" factors and ratings "method" factors. Quellmalz and Capell (1979) used multitrait-multimethod confirmatory factor analyses to examine discriminant validity of subscales generated by analytic scoring rubrics and the comparative information yield of alternative response modes for writing assessment (i.e., essay, paragraph and selected response). Their results indicated relatively high intercorrelations among subscale content factors, as well as a general tendency for the shorter assessment modes to generate less pure indicators of the subscale factors.

If non-native English speakers' English writing is easily distinguished from that of native speakers on the dimension of mechanics, and if such group differences contaminate other ratings assigned to non-native speakers, a straightforward form of bias may be present. Ratings on other dimensions will be systematically depressed, and the diagnostic utility of the writing appraisal undermined. The present study was conducted to evaluate such bias in the context of variations of ethnicity of both students and raters, and of prompts. Additionally, the nature of the task presented to the students in order to get them to write an essay was varied systematically.

METHOD

Subjects

One hundred and thirty students from fifth- and sixth-grade monolingual English classrooms in a moderately sized California school district were involved in this

study as a normal part of their classroom activities. These students were not members of bilingual programs although some were involved in remedial "pull-out" instruction. Of the 116 students who provided complete essays, half were Hispanic-surnamed. Raters were four teachers hired during school vacation, of whom two were Hispanic and two non-Hispanic. These raters were from different school districts and had no other contact of any kind with the students in this sample.

Instruments

The study used a standardized writing task with two topics, and a modified scoring rubric, which will be explained below, which has been shown to have acceptable validity and reliability (Quellmalz & Capell, 1979). The packet containing the essay writing task consisted of a face sheet for student's name and date, followed by two prompts and two lined response pages, totalling five pieces of paper per handout. The prompts involved two topics, one a main street of a town and the other a robot. Order of presentation of the prompts, and whether the prompt was written or pictorial, was controlled for every participant. Written prompts involved five lines of typewritten text, while picture prompts involved a lead sentence and a full-page line drawing of the children's topics by a graduate student artist. In both situations, the text concluded with the request that the student write a paragraph about the topic presented. No other information was made available to the student.

The raters reviewed these essays using the Center for the Study of Evaluation's Factual Narrative scoring rubric, consisting of four primary subscales -- General Impression, Focus and Organization, Support, and Grammar and Mechanics. Each of these was evaluated on a six-point scale, ranging from clear mastery of the assignment to clear failure. For each of the six values on each of the four scales, extensive guidelines for scoring were provided. General Impression rating of the

essay is formed by considering all aspects of the effectiveness of composition, including the remaining three rating criteria. The Focus and Organization subscale handles such issues as logical progression, transitions, and topic development. The Support subscale rates the use of specific supporting statements and details. The Grammar and Mechanics subscale is used to evaluate the essay's sentence construction, word usage, spelling and punctuation. In addition to an overall rating from this last subscale, the extent of errors of each of the four areas of Mechanics noted above is rated separately. The instructions of the CSE scoring rubric make explicit that raters using factual scoring will likely find that some qualities of an essay cannot be considered separate from others, but it is also quite direct in indicating how any particular rating is to correspond to the annotation supplied in the guidelines.

Procedure

Each child received one essay packet containing two essay prompts -- one pictorial and the other written -- and ruled pages for the child's essays. The package of essay prompts was administered in a single half-hour sitting by the children's classroom teachers, and essays were collected and sent directly for rating without further intervention in the classroom.

Each of the raters was given every essay packet in random order, but without the face sheet and thus without identification of the name or ethnic background of the student writers. Following five days of training and pilot testing on use of the CSE rating scales, the four raters completed scoring of the 116 essay packages which were complete and legible over a seven day period. The resulting 32 ratings for each essay (four raters x eight subscales) were then analyzed by a three factor analysis of variance (student ethnicity x rater ethnicity x prompt modality) with repeats on the second two factors (Winer, 1962) separately for each subscale. Also collected from school district records were subtest totals on the

Comprehensive Tests of Basic Skills (CTBS), administered as part of the regular testing program by the school district, for all students involved in the study. These scores allowed the investigation of possible relationships between the measures of writing capability and four aspects of students' intellectual capacity-- vocabulary, passage comprehension, language mechanics and expression.

RESULTS AND DISCUSSION

Only essays with complete ratings were considered in the analysis; complete data were available for the four primary subscales for 100 essays, and for the four detail subscales for 74 essays. Average rater agreement across all subscales was high for the two Hispanic raters (92.15%) and moderately good for the non-Hispanic raters (85.46%). When all four raters were compared, average agreement on the subscales was good (81.15%). These values were considered as acceptable evidence that the training of the raters had been satisfactory. To minimize potential confounding from differences between the two topics, all scores were then standardized within topic before further analysis.

On the General Impression subscale, the interaction between student ethnicity (Hispanic or non-Hispanic) and rater ethnicity (Hispanic or non-Hispanic) was significant ($F_{1,98}$=6.51, MSerror = 13.37, $p<.01$). While the non-Hispanic student essays received about the same General Impression scores from Hispanic raters as the Hispanic student essays, the non-Hispanic raters significantly favored the non-Hispanic student essays. No other main effect or interaction was significant for this subscale. The interaction between student ethnicity and rater ethnicity was also found on the Support subscale ($F_{1,98}$=4.02, MSerror = 31.48, $p<.05$), and on the Mechanics subscale ($F_{1,98}$=7.18, MSerror = 36.42, $p<.01$). On the Support subscale, the non-Hispanic student essays were again significantly favored by the non-Hispanic raters. However, on the Mechanics subscale, the non-Hispanic raters judged both student groups alike while the Hispanic raters gave the essays of the

non-Hispanic students significantly lower scores.

For the Focus subscale, a main effect of rater ethnicity ($F_{.,98}$=11.82, MSerrer = 16.62, p<.001) and an interaction between rater ethnicity and prompt mode (picture prompt or written prompt) ($F_{1,98}$= 6.41, MSerror = 19.01, p<.01) were found. In addition to the rater ethnicity by student ethnicity interactions, the Support subscale yielded only a main effect of prompt modality ($F_{1,98}$= 10.43, MSerror = 68.17, p<.001), and the Mechanics subscale yielded only a main effect of rater ethnicity ($F_{.,98}$= 13.45, MSerror = 36.42, p<.001). On the detail subscales of Mechanics, only one effect emerged as significant: rater ethnicity as a factor in Usage ratings ($F_{1,73}$= 41.01, MSerror = 47.01, p<.001). No other detail subscale showed any significant main effect or interaction. Table 1 summarizes the findings across the four primary and the usage detail subscales by main effect and interactions, and the results of post-hoc analyses.

When performance scores on the CTBS were compared, neither the Hispanic nor non-Hispanic students emerged as significantly more capable on any subscale than the others. The results of the correlational study between student essay ratings and the four selected scale scores from the CTBS can be summarized rapidly. Not a single significant correlation appeared between any rating subscale and any CTBS scale for this sample. Thus there appears to be no intrinsically overlapping information between writing performance as judged on CSE's Factual Narrative rubric and a sample of academic performance as judged on a multiple-choice examination.

The most important finding, repeated across three of the subscales, is that the student ethnicity and rater ethnicity factors interact frequently and substantively in the appraisal of students' written essays. Additionally, rater ethnicity alone is also a significant factor in the ratings. These results point to three conclusions. First, the evaluation of prose writing seems to be systematically

Table 1

Summary of Statistically Significant ($p<.05$) Effects

Subscale:	General Impression 100	Focus and Organization 100	Support 100	Mechanics 100	Usage Detail[1] 74
Main Effects					
Student Ethnicity	--	--	--	--	--
Rater Ethnicity	--	*[2]	--	*[2]	*[2]
Prompt	--	--	*[3]	--	--
Interactions					
Student x Rater	*[4]	--	*[4]	*[5]	--
Student x Prompt	--	--	--	--	--
Rater x Prompt	--	*[6]	--	--	--
Student x Rater x Prompt	--	--	--	--	--

[1]Remaining detail subscales show no significant effects.

[2]Hispanic raters elevated relative to non-Hispanic raters.

[3]Picture prompt elevated relative to written prompt.

[4]Non-Hispanic raters + non-Hispanic student essays elevated relative to other combinations.

[5]Hispanic raters + non-hispanic student essays depressed relative to other combinations.

[6]Non-Hispanic raters + Hispanic student essays elevated relative to other combinations.

affected by factors which reflect different cultural backgrounds. It is important to note that this effect does not emerge when essays are grouped solely by student ethnicity; rather, the students of one or the other backgrounds were often judged differently by raters who share that background than by raters who do not. Second, these factors include (but are not limited to) a match or mismatch between raters' and writers' preferred language styles, and to some extent the nature of the stimulus used to initiate the writing sample. Note, however, that the three factor interaction between student ethnicity, rater ethnicity and type of prompt was not observed for any of the subscales used. Third, the phenomenon of systematic matching or mismatching of preferences and styles occurs despite the fact that the evaluative scheme used is one with a high degree of objectivity, which would be expected to minimize such matching relative to more subjective rating scheme. the nature of the judgment task is referenced point-for-point by the CSE scoring rubric and thus no scale-free or endpoint-only continuum judgments were involved. Additionally, because raters were blind not only to the names and ethnicities of the essay writers, but to the study's hypotheses and the proportional representation of ethnicities within the sample, whatever matching occured most likely stems from recognition of and preference for certain subtle aspects of writing styles.

Some limitations of the present study deserve attention. There are many possible secondary analyses of writing style, process and content which have not been pursued here. No information about essay complexity or other linguistic patterns is available from the present analysis. How creative, stereotyped, or bizarre the particular essay is goes unremarked in the CSE scoring system. The isolation of exact details within essay content or specific preferences of individual raters was not within the purview of this investigation. Moreover, there is a small possibility that systematic differences in handwriting mastery contributed to the recognizability of student ethnicity and thus to the ratings

given, but this was not examined directly. None of these considerations is seen as critical to the interpretation of the results presented above, in particular because the expected outcome of the analyses of variance in such instance would necessarily be a main effect due to student ethnicity alone or a three-way interaction between student ethnicity, rater ethnicity and prompt modality. None of these effects emerged in the present study, but rather a pattern of findings which strongly suggests that some complex form of bias is at work.

Bias in judgment is a phenomenon which obtains under a variety of circumstances, some of which are intrinsic in the testing and evaluation process. The present findings indicate that extrinsic factors must also be considered. In the case of judgment of essays, where essay content has virtually limitless possibilities and appraisal is of necessity at least partially subjective, the opportunity for unintentional bias seems more likely. For the teacher or essay test administrator seeking to limit bias to the absolute minimum, the mandate is: those who are to perform the rating of the essays must be matched for appropriate backgrounds of the students who write the essays and are judged.

REFERENCES

Bikson, T.K. Do they talk the same language? Lexical interface and ethnicity. Santa Monica, Rand Corporation, 1977, mimeo.

Caldwell, F.F.,& Mowry, M.D. The essay versus the objective examination as measures of achievement in bilingual children. Journal of Educational Psychology, 1933, 24, 696-702.

Cervantes, R.A. Self-concept, locus of control, and achievement in Mexican-American pupils. Unpublished doctoral dissertation, San Francisco: Union Graduate School-West, 1975

Follman, J.C., & Anderson, J.A. An investigation of the reliability of five procedures for grading English theses. Research in the Teaching of English, 1967, 1, 190-200.

Freedman, S. Why do teachers give the grades they do? College Composition and Communication, 1979, 30, 161-164.

Garcia, R. A linguistic frame of reference for critiquing Chicano compositions, College English, 1975/76, 37, 184-188.

Hartwell, P. Writers as readers, Paper presented at the annual meeting of the Conference on College Composition and Communication, Dallas, 1981. ERIC Document Number ED199701.

Jaeger, R.M. & Freijo, T.D. Race and sex as concomitants of composite halo in teachers' evaluative ratings of pupils. Journal of Educational Psychology, 1975, 67, 226-237.

Lambert, W.E., Gardner, R.C., Barik, H.C, & Tunstall, K. Attitudinal and Cognitive aspects of intensive study of a second language. Journal of Abnormal and Social Psychology, 1963, 66, 358-368.

Lay, N.D.S. Chinese language interference in written English. Journal of Basic Writing, 1978, 1 50-61.

Olson, D.R. From utterance to text: The bias of language in speech and writing. Harvard Educational Review, 1977, 47, 257-281.

Quellmalz, E. & Capell, F. Defining writing domains: Effects of discourse and response mode. Los Angeles: Center for the Study of Evaluation, Grant Number B-NIE-G-78-0213.

Rhodes-Hoover, M, & Politzer, R.L. Bias in composition tests with suggestions for a culturally appropriate assessment technique. Paper presented at the National Institute of Education Writing Conference, Washington, D.C., 1977.

Rissel, D. Implications of differences in the organizations of a lexical domain in Spanish and English bilinguals. Bilingual Review, 1978, 3, 29-34.

Rizzo, B. & Villafane, S. Spanish language influences on written English. Journal of Basic Writing, 1978, 1, 62-71.

Rodrigues, R. A statistical study of the English syntax of bilingual Mexican-American and monolingual Anglo-American students. Bilingual Review, 1978, 3, 205-211.

Smith, F. Spoken and written language. In E.H. Lenneberg & E. Lenneberg, (Eds.) Foundations of language development, a multidisciplinary approach, New York: Academic Press, 1975

Stahl, A. The structure of children's compositions: Developmental and ethnic differences. Research in the Teaching of English, 1977, 11, 156-163.

Sternglass, M.S. Assessing reading, writing and reasoning. College English, 1981, 43, 269-275.

White, E.M. & Thomas, L.L. Racial minorities and writing skills assessment in the California State University and Colleges. College English, 1981, 43, 276-283.

Winer, B.J. Statistical principles in experimental design. New York, McGraw-Hill, 1962.